

Supplementary online materials for:

Evidence for APOBEC3B mutagenesis in multiple human cancers

Michael B. Burns¹⁻⁴, Nuri A. Temiz^{1,2} & Reuben S. Harris^{1-4, #}

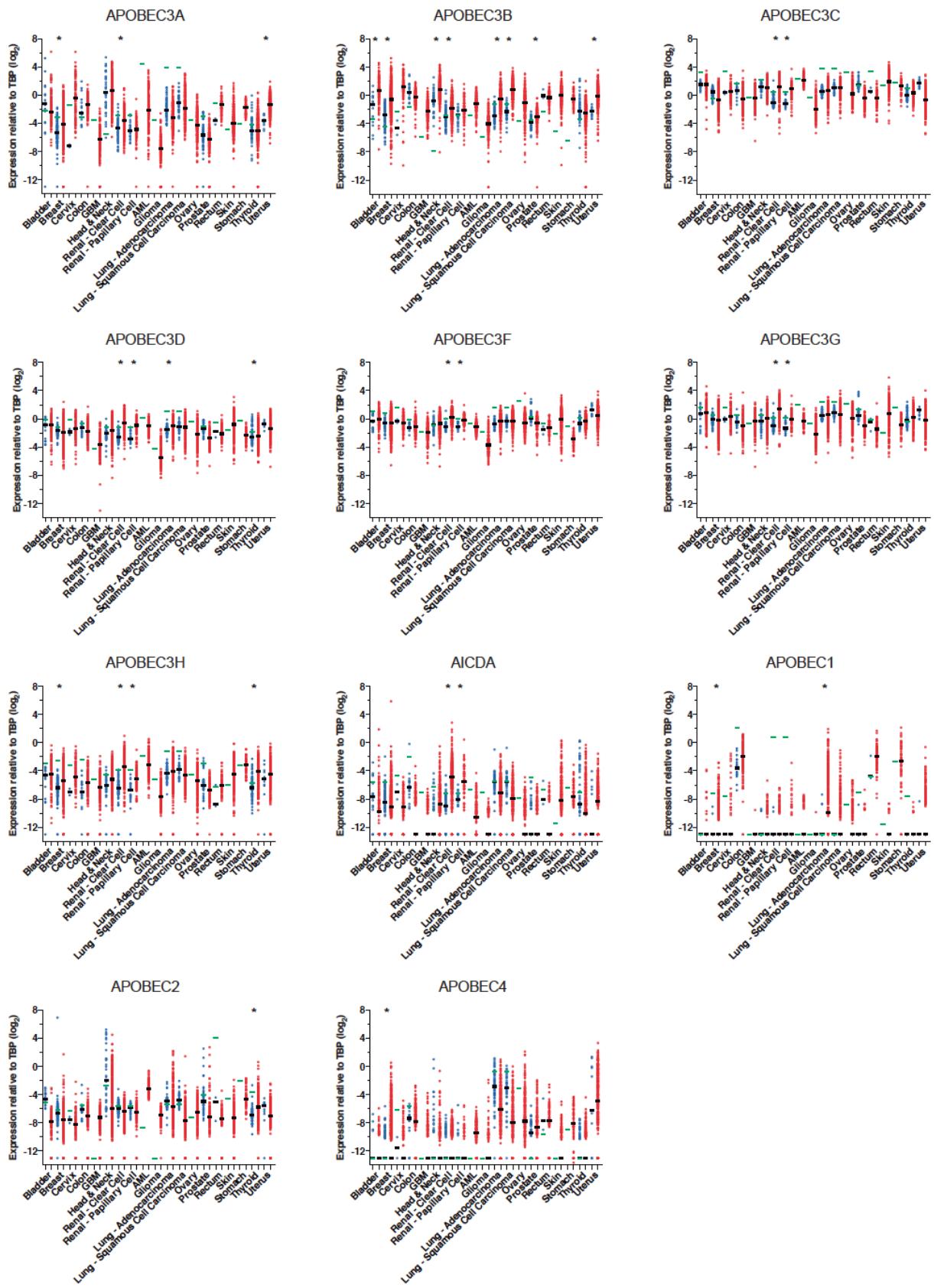
¹Biochemistry, Molecular Biology and Biophysics Department, ²Masonic Cancer Center,

³Institute for Molecular Virology, ⁴Center for Genome Engineering, University of Minnesota,

Minneapolis, MN 55455, USA

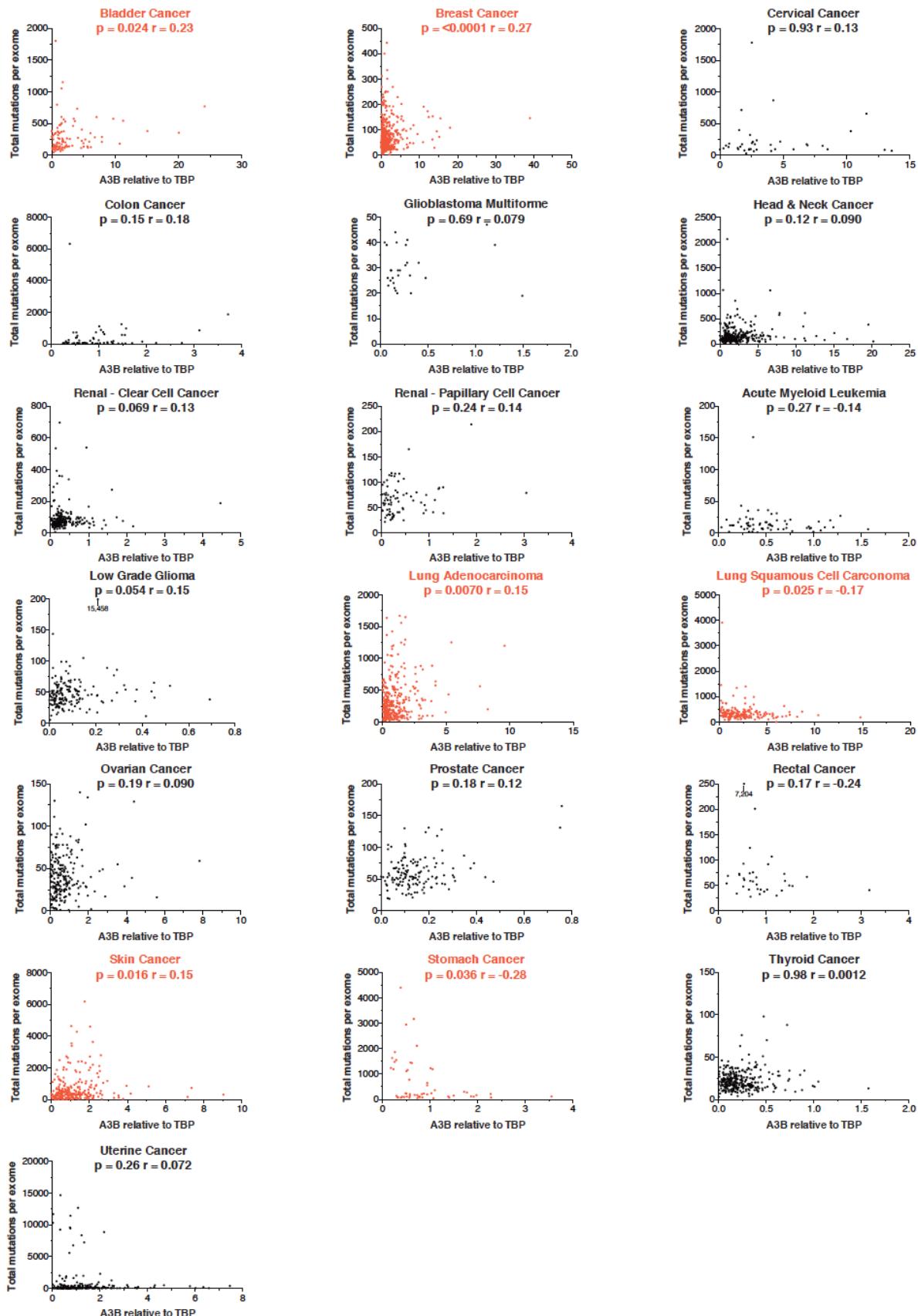
[#]Correspondence to R.S.H. (rsh@umn.edu).

This section contains Supplementary Figures 1-4 and Supplementary Tables 1-3.

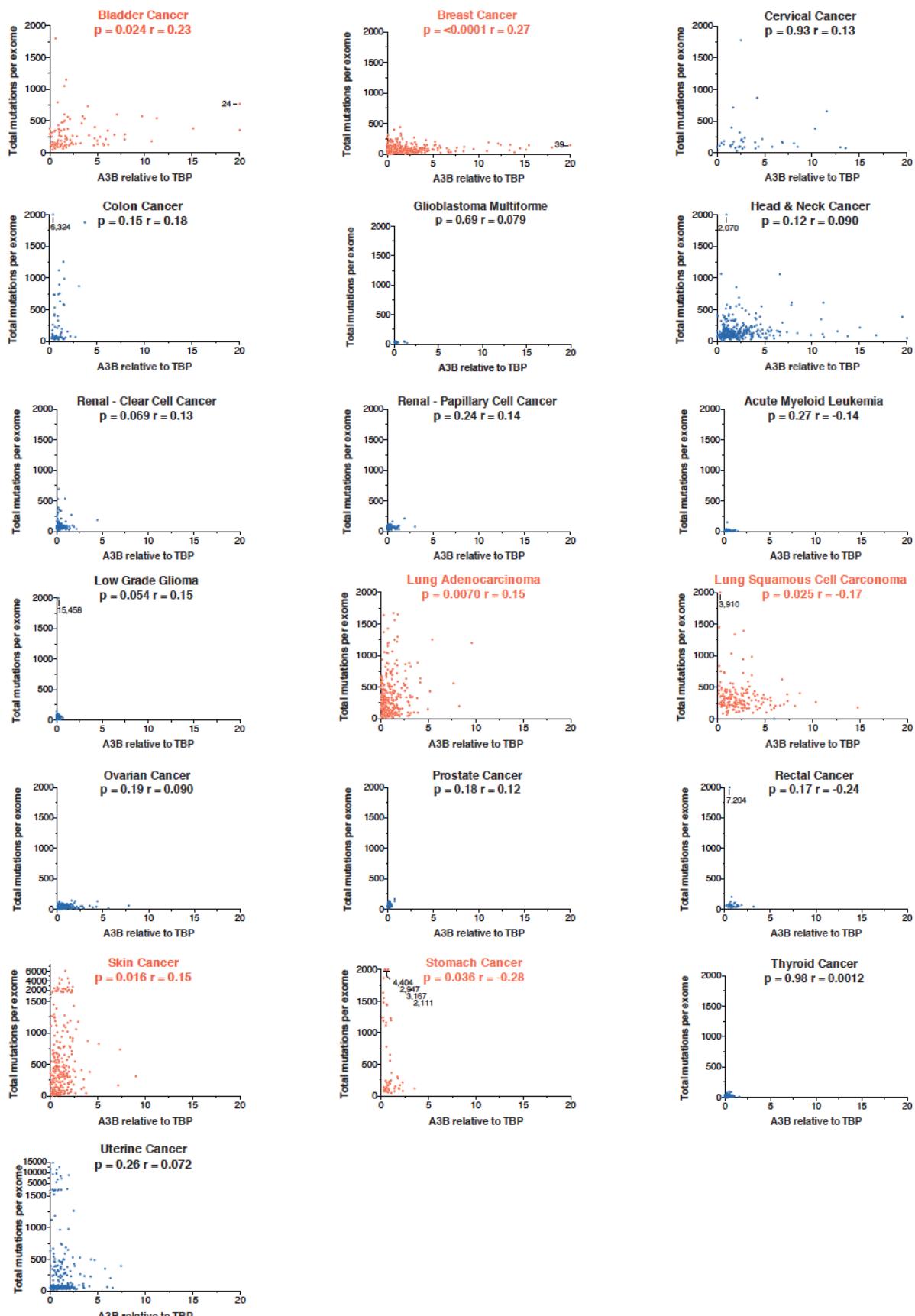


Supplementary Fig. 1. *APOBEC* family member mRNA expression levels for all 19 cancers analyzed here. See next page for full legend.

Supplementary Fig. 1. *APOBEC* family member mRNA expression levels for all 19 cancers analyzed here. RNAseq and RT-qPCR data for expression of the indicated *APOBEC* family member genes relative to the housekeeping gene, *TBP*. Each data point represents one tumor (red symbol) or normal (blue symbol) sample, and the Y-axis is log-transformed for better data visualization. Black horizontal lines indicate the median *APOBEC/TBP* value for each cancer or normal data set (**Table 1** and **Supplementary Table 1**). Green horizontal lines indicate the *APOBEC/TBP* value determined by RT-qPCR. Asterisks indicate significant upregulation of the indicated gene in the tumor relative to the corresponding normal tissues ($p < 0.0001$ by Mann-Whitney U-test). *APOBEC3B* expression data are reproduced from **Fig. 1** for comparison with other family members. The positive expression correlations in the two types of renal tumors for nearly all *APOBEC* family members cannot be explained at this time. The positive association of *APOBEC3A* in breast and bladder cancer may be due to infiltrating macrophages, as this mRNA is only expressed in myeloid lineage cell types and is not present in breast cancer cell lines (refs. 33 & 35). The positive correlations for *APOBEC3D* in lung adenocarcinoma and thyroid cancers barely reach significance. The positive correlations for *APOBEC3H*, *APOBEC1*, and *APOBEC4* in breast cancer were not observed previously by RT-qPCR in tumors with patient-matched normal tissues as controls (ref. 33). The positive correlations for *APOBEC3H* and *APOBEC2* in thyroid cancer and *APOBEC1* in lung adenocarcinoma are not explainable at this time and could be interesting subjects for further work. P-values for negative or insignificant associations are not indicated in this figure. Overall, although these data indicate that *APOBEC3B* is the most abundantly upregulated *APOBEC* family member across the many different cancers, these data are only one line of evidence suggesting a role in cancer and they must be interpreted in alongside other analyses presented here and in prior literature, which impose strong biochemical, genetic, and cellular constraints on what is and is not possible or plausible (see **Results** and **Discussion**).



Supplementary Fig. 2a. Correlations between total mutation loads and *APOBEC3B* expression levels.
See page 6 for full legend.

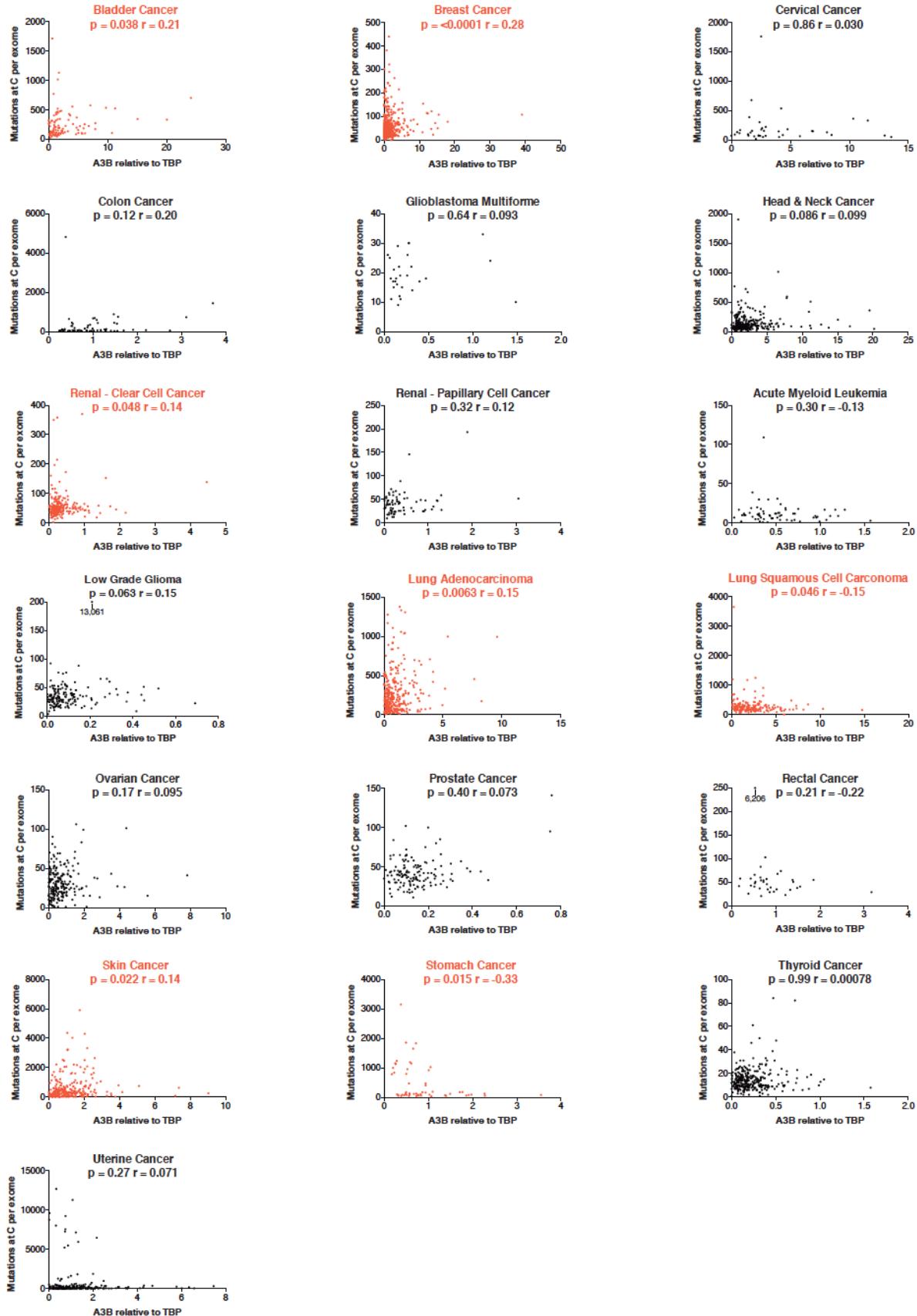


Supplementary Fig. 2b. Correlations between total mutation loads and *APOBEC3B* expression levels.
See page 6 for full legend.

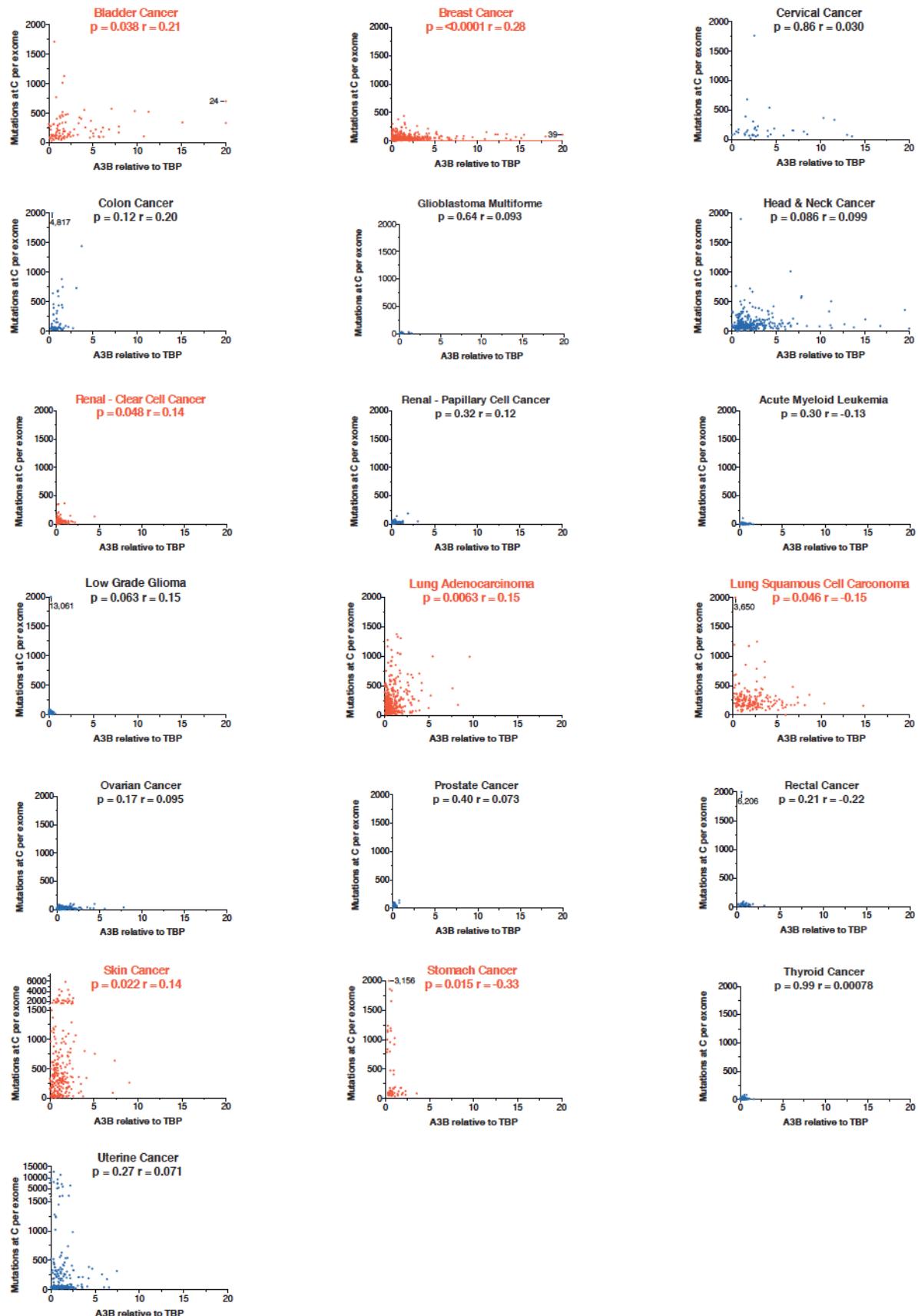
Supplementary Fig. 2. Correlations between total mutation loads and *APOBEC3B* expression levels.

(a) Total exonic mutation loads plotted against *APOBEC3B/TBP* expression levels for each of the 19 tumor types analyzed here. P and r-values are from Spearman's correlation. Data sets with p-values less than or equal to 0.05 are highlighted in red. The high variability in mutation loads amongst each tumor type is due to the stochastic nature of the underlying mutational processes, different tumor ages, differential repair capacities, selection bottlenecks, chemotherapeutic drug exposures, etc.

(b) The same data as in panel (a) but projected onto fixed axes to facilitate comparison between tumor types.



Supplementary Fig. 3a. Correlations between C/G-specific mutation counts and *APOBEC3B* expression levels. See page 9 for full legend.

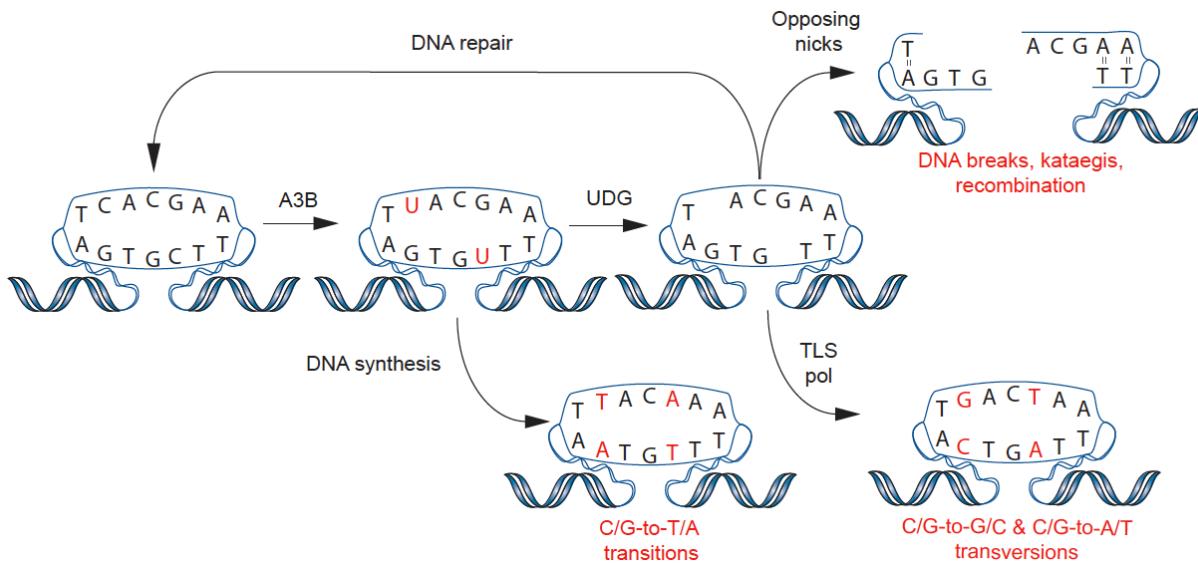


Supplementary Fig. 3b. Correlations between C/G-specific mutation counts and *APOBEC3B* expression levels. See page 9 for full legend.

Supplementary Fig. 3. Correlations between C/G-specific mutation counts and *APOBEC3B* expression levels.

(a) Exonic C/G mutation counts plotted against *APOBEC3B/TBP* expression levels for each of the 19 tumor types analyzed here. P and r-values are from Spearman's correlation. Data sets with p-values less than or equal to 0.05 are highlighted in red. The high variability in mutation loads among each tumor type is due to the stochastic nature of the underlying mutational processes, different tumor ages, differential repair capacities, selection bottlenecks, chemotherapeutic drug exposures, etc.

(b) The same data as in panel (a) but projected onto fixed axes to facilitate comparison between tumor types.



Supplementary Fig. 4. Model for APOBEC3B-induced mutagenesis in cancer.

APOBEC3B deaminates genomic cytosines in preferred contexts resulting in uracils. DNA repair by uracil DNA glycosylase (UDG) and canonical base excision repair may correct many lesions. C-to-T transitions may result from DNA synthesis templated directly by genomic uracils or from DNA synthesis to bypass abasic sites (following established ‘A-rule’, not shown). C-to-G and C-to-A transversions may result during bypass of template abasic sites by a translesion synthesis DNA polymerase (TLS pol). Abasic sites may be further processed by a base excision repair endonuclease (APEX, not shown) into nicks, which can lead to single- and double-stranded DNA breaks, to exposed single-stranded DNA and kataegis events, as well as to recombination and larger-scale genomic aberrations such as translocations. Model adapted from Ref. 33.

Supplementary Table 1. Summary statistics for the normal control samples in this study.

Tumor Type	TCGA ID	A3B expression in normal controls ¹			A3B expression in normal controls ²
		n	Range	Median	
Low Grade Glioma	LGG	n.a.	n.a.	n.a.	0.016
Prostate adenocarcinoma	PRAD	44	0.017 - 0.21	0.41	0.090
Thyroid carcinoma	THCA	58	0.0058 - 5.1	1.0	0.10
Glioblastoma multiforme	GBM	n.a.	n.a.	n.a.	0.016
Kidney renal papillary cell carcinoma	KIRP	25	0.029 - 0.43	0.10	0.14
Kidney renal clear cell carcinoma	KIRC	71	0.024 - 1.7	0.25	0.14
Acute myeloid leukemia	LAML	n.a.	n.a.	n.a.	0.092
Ovarian serous cystadenocarcinoma	OV	n.a.	n.a.	n.a.	0.080
Breast invasive carcinoma	BRCA	107	0.0081 - 0.69	0.15	0.048
Stomach adenocarcinoma	STAD	n.a.	n.a.	n.a.	0.012
Lung adenocarcinoma	LUAD	57	0.037 - 0.89	0.16	0.44
Rectum adenocarcinoma	READ	3	0.78 - 1.8	0.54	0.21
Colon adenocarcinoma	COAD	18	0.46 - 7.7	2.0	0.34
Uterine corpus endometrioid carcinoma	UCEC	11	0.10 - 0.42	0.10	n.a.
Skin cutaneous melanoma	SKCM	n.a.	n.a.	n.a.	0.030
Bladder urotheilal carcinoma	BLCA	16	0.014 - 2.6	0.66	0.10
Head & neck squamous cell carcinoma	HNSC	37	0.049 - 5.9	1.0	0.0042
Lung squamous cell carcinoma	LUSC	35	0.027 - 0.77	0.16	0.44
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	2	0.021 - 0.085	0.099	0.20

¹A3B expression values relative to those of the housekeeping gene *TBP*, determined by RNAseq.

²A3B expression values relative to those of the housekeeping gene *TBP*, determined by qPCR.

Supplementary Table 2. Euclidean distances between each tumor type and the signature of recombinant APOBEC3B (recA3B).

	recA3B	BLCA	BRCA	CESC	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
recA3B	-	0.180	0.178	0.190	0.328	0.241	0.162	0.213	0.179	0.299	0.302	0.179	0.154	0.202	0.271	0.311	0.320	0.337	0.220	0.278
BLCA	0.180	-	0.123	0.040	0.317	0.221	0.102	0.219	0.160	0.288	0.299	0.202	0.173	0.203	0.258	0.295	0.293	0.316	0.182	0.300
BRCA	0.178	0.123	-	0.135	0.211	0.132	0.036	0.115	0.064	0.175	0.197	0.134	0.111	0.092	0.140	0.189	0.295	0.210	0.078	0.217
CESC	0.190	0.040	0.135	-	0.322	0.236	0.116	0.235	0.176	0.296	0.308	0.221	0.189	0.218	0.266	0.299	0.312	0.319	0.193	0.309
COAD	0.328	0.317	0.211	0.322	-	0.217	0.233	0.186	0.203	0.100	0.093	0.260	0.256	0.193	0.099	0.112	0.394	0.057	0.167	0.171
GBM	0.241	0.221	0.132	0.236	0.217	-	0.147	0.139	0.133	0.167	0.205	0.167	0.165	0.110	0.142	0.195	0.312	0.214	0.128	0.239
HNSC	0.162	0.102	0.036	0.116	0.233	0.147	-	0.126	0.071	0.199	0.215	0.125	0.097	0.108	0.165	0.215	0.289	0.235	0.097	0.230
KIRC	0.213	0.219	0.115	0.235	0.186	0.139	0.126	-	0.067	0.151	0.159	0.108	0.109	0.060	0.118	0.197	0.312	0.202	0.086	0.199
KIRP	0.179	0.160	0.064	0.176	0.203	0.133	0.071	0.067	-	0.169	0.177	0.110	0.096	0.065	0.133	0.203	0.288	0.214	0.065	0.203
LAML	0.299	0.288	0.175	0.296	0.100	0.167	0.199	0.151	0.169	-	0.138	0.223	0.220	0.148	0.059	0.098	0.363	0.087	0.127	0.207
LGG	0.302	0.299	0.197	0.308	0.093	0.205	0.215	0.159	0.177	0.138	-	0.225	0.225	0.166	0.124	0.160	0.374	0.131	0.159	0.140
LUAD	0.179	0.202	0.134	0.221	0.260	0.167	0.125	0.108	0.110	0.223	0.225	-	0.045	0.102	0.192	0.253	0.322	0.273	0.154	0.243
LUSC	0.154	0.173	0.111	0.189	0.256	0.165	0.097	0.109	0.096	0.220	0.225	0.045	-	0.099	0.187	0.246	0.316	0.267	0.141	0.241
OV	0.202	0.203	0.092	0.218	0.193	0.110	0.108	0.060	0.065	0.148	0.166	0.102	0.099	-	0.113	0.183	0.311	0.199	0.082	0.202
PRAD	0.271	0.258	0.140	0.266	0.099	0.142	0.165	0.118	0.133	0.059	0.124	0.192	0.187	0.113	-	0.099	0.349	0.096	0.097	0.187
READ	0.311	0.295	0.189	0.299	0.112	0.195	0.215	0.197	0.203	0.098	0.160	0.253	0.246	0.183	0.099	-	0.382	0.080	0.165	0.192
SKCM	0.320	0.293	0.295	0.312	0.394	0.312	0.289	0.312	0.288	0.363	0.374	0.322	0.316	0.311	0.349	0.382	-	0.398	0.291	0.368
STAD	0.337	0.316	0.210	0.319	0.057	0.214	0.235	0.202	0.214	0.087	0.131	0.273	0.267	0.199	0.096	0.080	0.398	-	0.171	0.202
THCA	0.220	0.182	0.078	0.193	0.167	0.128	0.097	0.086	0.065	0.127	0.159	0.154	0.141	0.082	0.097	0.165	0.291	0.171	-	0.202
UCEC	0.278	0.300	0.217	0.309	0.171	0.239	0.230	0.199	0.203	0.207	0.140	0.243	0.241	0.202	0.187	0.192	0.368	0.202	0.202	-

Supplementary Table 3. Description of the mutation subset analysed in this study.

Tumor Type	TCGA ID	Number of tumors	Total mutations	Filtered mutations	Percent of mutations filtered (non-SNP)
Low Grade Glioma	LGG	170	24650	1213	5%
Prostate adenocarcinoma	PRAD	150	9784	881	9%
Thyroid carcinoma	THCA	326	12143	4826	40%
Glioblastoma multiforme	GBM	167	5862	146	2%
Kidney renal papillary cell carcinoma	KIRP	100	8068	1167	14%
Kidney renal clear cell carcinoma	KIRC	244	33280	10811	32%
Acute myeloid leukemia	LAML	74	1368	137	10%
Ovarian serous cystadenocarcinoma	OV	469	28049	2227	8%
Breast invasive carcinoma	BRCA	777	52160	6290	12%
Stomach adenocarcinoma	STAD	156	100913	14899	15%
Lung adenocarcinoma	LUAD	392	152307	13269	9%
Rectum adenocarcinoma	READ	88	21199	1181	6%
Colon adenocarcinoma	COAD	266	148114	18503	12%
Uterine corpus endometrioid carcinoma	UCEC	248	184829	5719	3%
Skin cutaneous melanoma	SKCM	255	186839	9207	5%
Bladder urotheelial carcinoma	BLCA	99	30801	1948	6%
Head & neck squamous cell carcinoma	HNSC	306	63508	8282	13%
Lung squamous cell carcinoma	LUSC	177	65306	967	1%
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	39	10021	936	9%